UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Miloš Milaković

SEKVENCIRANJE ANTIBIOTIKA - ELEKTRONSKA VERZIJA

master rad

Beograd, 2025.

Mentor:

dr Jovana Kovačević, vandredni profesor
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

dr Mirjana Maljković Ružičić, docent
Univerzitet u Beogradu, Matematički fakultet

dr Nevena Ćirić, asistent
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

Najvoljenijima

Naslov master rada: Sekvenciranje antibiotika - Elektronska verzija

Rezime: Ovaj rad predstavlja elektronsku lekciju posvećenu sekvenciranju antibiotika, sa fokusom na interaktivno upoznavanje sa različitim algoritamskim pristupima. Lekcija obuhvata teorijsko objašnjenje i vizuelizaciju algoritama kao što su gruba sila, *Branch and Bound*, *Leaderboard* algoritam, spektralna konvolucija i *DeepNovo* sekvenciranje. Korisnicima je omogućeno da prate izvršavanje algoritama korak po korak, sa opcijama pauziranja i ponavljanja, čime se olakšava razumevanje kompleksnih procesa sekvenciranja. Ova interaktivna platforma može služiti kao edukativni alat za studente i nastavnike u oblasti bioinformatike i hemijske analize antibiotika.

Ključne reči: sekvenciranje, antibiotici, algoritmi, računarstvo, aminokiseline, maseni spektrometar, gruba sila, *Branch and Bound*, *Leaderboard*, *DeepNovo*, spektralna konvolucija

Sadržaj

Glava 1

Uvod

Sekvenciranje antibiotika predstavlja ključni proces u savremenoj bioinformatici i farmaceutskoj industriji. Antibiotici su supstance koje inhibiraju rast mikroorganizama ili ih uništavaju, što ih čini neophodnim u lečenju bakterijskih infekcija. Većina klinički relevantnih antibiotika su peptidne prirode, što znači da su oni zapravo kratki proteini odnosno da su sastavljeni od lanaca aminokiselina - osnovnih gradivnih jedinica proteina.

U prirodi postoji 20 standardnih aminokiselina (slika 1.1) koje se kombinuju u različitim sekvencama da formiraju peptide i proteine. U kontekstu antibiotika, specifičan redosled aminokiselina je kritičan jer određuje:

- Trodimenzionalnu strukturu molekula
- Biološku aktivnost i mehanizam dejstva

1.1 Značaj sekvenciranja antibiotika

Precizno određivanje sekvence aminokiselina koje čine antibiotik ima brojne praktične primene:

- **Proizvodnja generičkih lekova**: Neophodno za reprodukciju postojećih antibiotika
- **Istraživanje otpornosti**: Pomaže u razumevanju mehanizama bakterijske otpornosti
- **Kontrola kvaliteta**: Verifikacija strukture proizvedenih antibiotika

- **Otkrivanje novih antibiotika**: Identifikacija nepoznatih prirodnih jedinjenja

Aminokiselina	**Skraćenica**	**Masa (Da)**
Glycine	G	57
Alanine	A	71
Serine	S	87
Proline	P	97
Valine	V	99
Threonine	T	101
Cysteine	C	103
Isoleucine	I	113
Leucine	L	113
Asparagine	N	114
Aspartic Acid	D	115
Lysine	K	128
Glutamine	Q	128
Glutamic Acid	E	129
Methionine	M	131
Histidine	H	137
Phenylalanine	F	147
Arginine	R	156
Tyrosine	Y	163
Tryptophan	W	186

Slika 1.1: Tabela masa aminokiselina izraženih u daltonima (Da)

1.2 Cilj rada

U ovom radu predstavljamo interaktivnu elektronsku platformu za sekvenciranje antibiotika koja kombinuje različite algoritamske pristupe sa vizuelizacijom u realnom vremenu. Platforma omogućava korisnicima da istraže proces sekvenciranja kroz interaktivne simulacije, čime se znatno olakšava razumevanje kompleksnih bioinformatičkih koncepata.

Glavni ciljevi ovog rada su:

- Pružiti pregled modernih algoritama za sekvenciranje antibiotika
- Razviti interaktivnu edukativnu platformu za vizuelizaciju sekvenciranja

Rad je posebno koristan studentima bioinformatike i molekularne biologije pružajući im alat za bolje razumevanje osnovnih principa masene spektrometrije i algoritama za rekonstrukciju peptidnih sekvenci.

Glava 2

Teorijske osnove

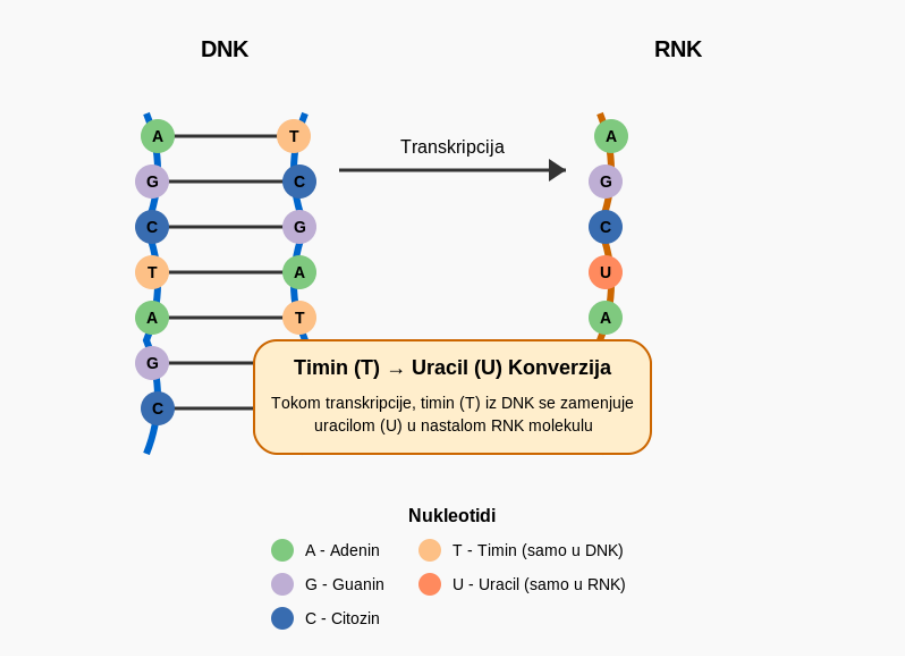
2.1 Centralna dogma molekularne biologije

Proces sekvenciranja antibiotika je fundamentalan u razumevanju kako su ovi molekuli proizvedeni od strane bakterija i kako se oni mogu sintetizovati ili modifikovani za primene u medicini. Antibiotici su često peptidi ali mnogo antibiotika, uglavnom neribozomalni peptidi (*non-ribosomal peptides - NRPs*), ne prati standardna pravila za sintezu proteina čime se otežava njihovo sekvenciranje [**?**].

DNK sadrži recept za kreiranje proteina. Odnosno, sastoji se od gena koji mogu biti uključeni i tada će se na osnovu njih kreirati proteini ili isključeni kada se oni neće koristiti za kreiranje proteina. Isključenost ili uključenost nekog gena zavisi od toga da li je potrebno da se kreira neki protein ili nije potrebno (npr. fotosinteza kod biljaka koja se obavlja samo preko dana).

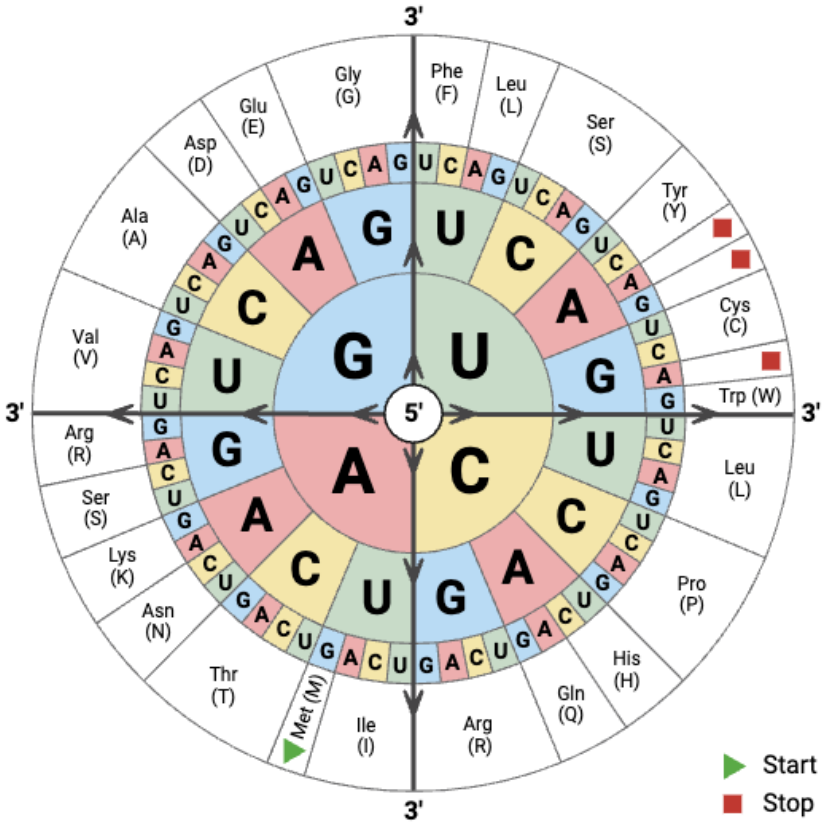
Tradicionalno, proteini prate Centralnu Dogmu Molekularne biologije, koja kaže da se DNK prvo prepisuje u RNK (Slika 2.1), a zatim se RNK prevodi u protein. Na slici 2.1 se može se primetiti da se DNK sastoji od 2 lanca koja su komplementarna. Enzim RNK polimeraza se kači na početak gena i kreće kroz gene gde razdvaja lanac i stvara prostor za prepisivanje DNK u RNK čime se dobija RNK.

Prilikom prevođenja RNK u protein potrebno je na osnovu nukleitoda odrediti koja je aminokiselina u pitanju. Organela ribozom je zadužena da odradi ovaj posao i pošto je potrebno na osnovu nukleotidne sekvence uniformno odrediti koja je aminokiselina u pitanju uzima se sekvenca od 3 nukleotida takođe poznata kao kodon. Pošto je uzeta sekvenca od 3 nukleotida ovo nam daje 64 različita kodona koja treba da se prevedu u 20 aminokiselina, da smo uzeli sekvencu od 2 nukleotida dobili bismo 16 različitih kombinacija čime ne bismo mogli da dobijemo sve aminokiseline.



Slika 2.1: Transkripcija DNK u RNK. Enzim RNK polimeraza (nije prikazan) čita DNK lanac i sintetiše komplementarni RNK lanac.

Na slici 2.2 može se videti kako se kodoni prevode u odgovarajuće aminokiseline. Postoje start i stop kodoni koji određuju početak odnosno kraj sekvence koja se prevodi u protein.

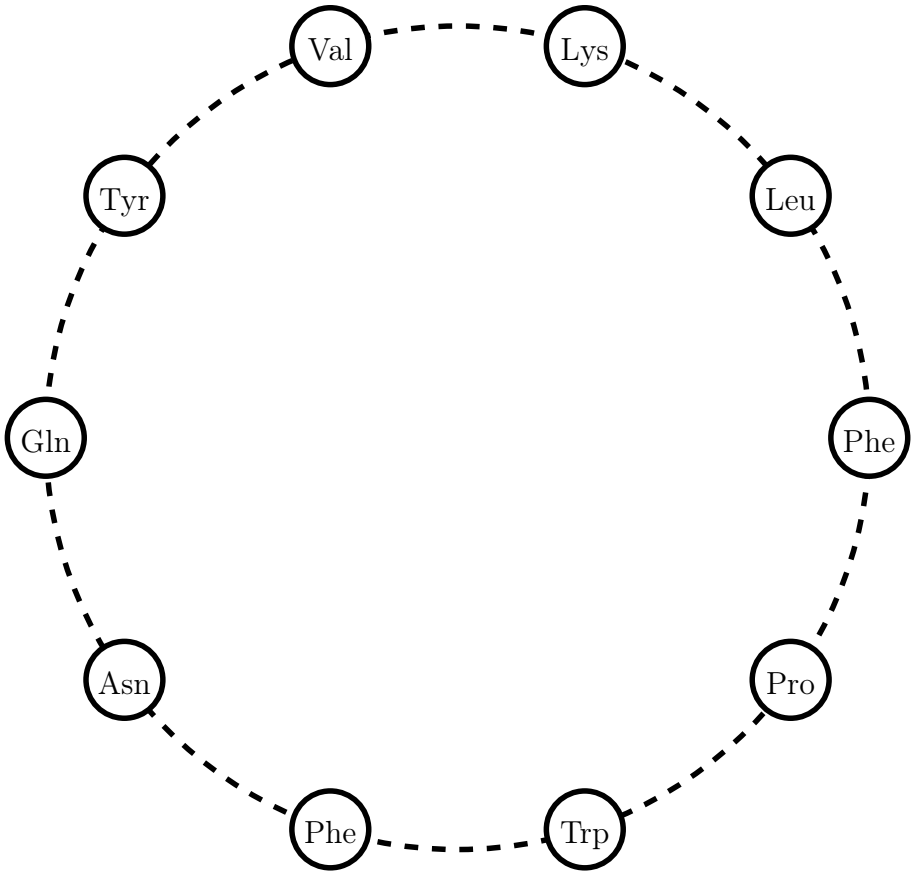


Slika 2.2: RNK kodonski točak prikazuje kako se sekvence od tri nukleotida (kodoni) prevode u aminokiseline. Svaki kodon se čita od centra ka spolja, a zeleni trougao označava start kodon (AUG) koji kodira metionin, dok crveni kvadrati označavaju stop kodone (UAA, UAG, UGA) koji određuju kraj sekvence koja se prevodi u protein. Preuzeto sa [?].

2.2 Odstupanje od centralne dogme

Tirocidin B1 je cikličan peptid dužine 10 (Slika 2.3), što znači da su prva i poslednja aminokiselina povezane i da samim tim postoji 10 njegovih različitih linearnih reprezentacija. Prateći centralnu dogmu i zaključka da se 1 kodon prevodi u 1 aminokiselinu, naučnici su probali da pronađu 10 kodona odnosno 30 nukleotida u genomu bakterije *Bacillus brevis* od koje nastaje ovaj antibiotik. Ovaj postupak je veoma dugotrajan obzirom da mora da se proveri više hiljada 30-grama koji mo-

gu da počnu bilo gde u genomu. Analiziranjem genoma utvrđeno je da ne postoji 30-gram koji se kodira u neki od 10 različitih reprenzatacija traženog antibiotika.



Slika 2.3: Struktura tirocidina B1, cikličnog peptida sastavljenog od 10 aminokiselina.

Dokazano je da Tirocidin B1 ne prati centralnu dogmu molekularne biologije i da postoje posebni enzimi koji su zaduženi za njihovo sintentisanje. Ovi enzimi se zovu NRP sintetaza. Ovi enzimi sadrže komplikovane module, koji govore koje aminokiseline učestvuju u sastavu proteina. U slučaju Tirocidina B1, enzim sadrži 10 modula i svaki od module kodira 1 aminokiselinu čime je određena struktura antibiotika.

Samim tim, pošto struktura proteina nije određena na osnovu genoma bakterije, metode za sekvencioniranje DNK ovde nisu od pomoći i potrebno je sekvencirati direktno sam peptid.

2.3 Maseni spektrometar

Maseni spektrometar [?] je moćan alat pomoću koga mogu da se odrede mase molekula, uključujući mase peptida i proteina. Omogućava naučnicima da odrede nepoznate komponente, saznaju strukturu molekula i analiziraju kompleksne uzorke. Maseni spektrometar radi tako što mu se da više uzoraka istog peptida a on napravi sve moguće podpeptide datog peptida i odredi njihove mase. U realnosti uzorak se pretvara u naelektrisane jone da bi na njih mogli da utiču električno i magnetno polje. Potom se joni dele na osnovu odnosa njihove mase i naelektrisanja i kao takvi se mere njihove vrednosti.

Masa se meri u daltonima (Da), pri čemu je 1 Da približno jednak masi protona/neutrona. Samim tim masa molekula je jednaka sumi masa protona/neutrona koji čine taj molekul. Mase aminokiselina su poznate i prikazane su na Slici 1.1. Može se primetiti da neke aminokiseline imaju istu masu, tako da 20 različitih aminokiselina ima 18 različitih masa.

Samim tim na osnovu poznatih masa aminokiselina možemo da odredimo da je masa tirocidina:

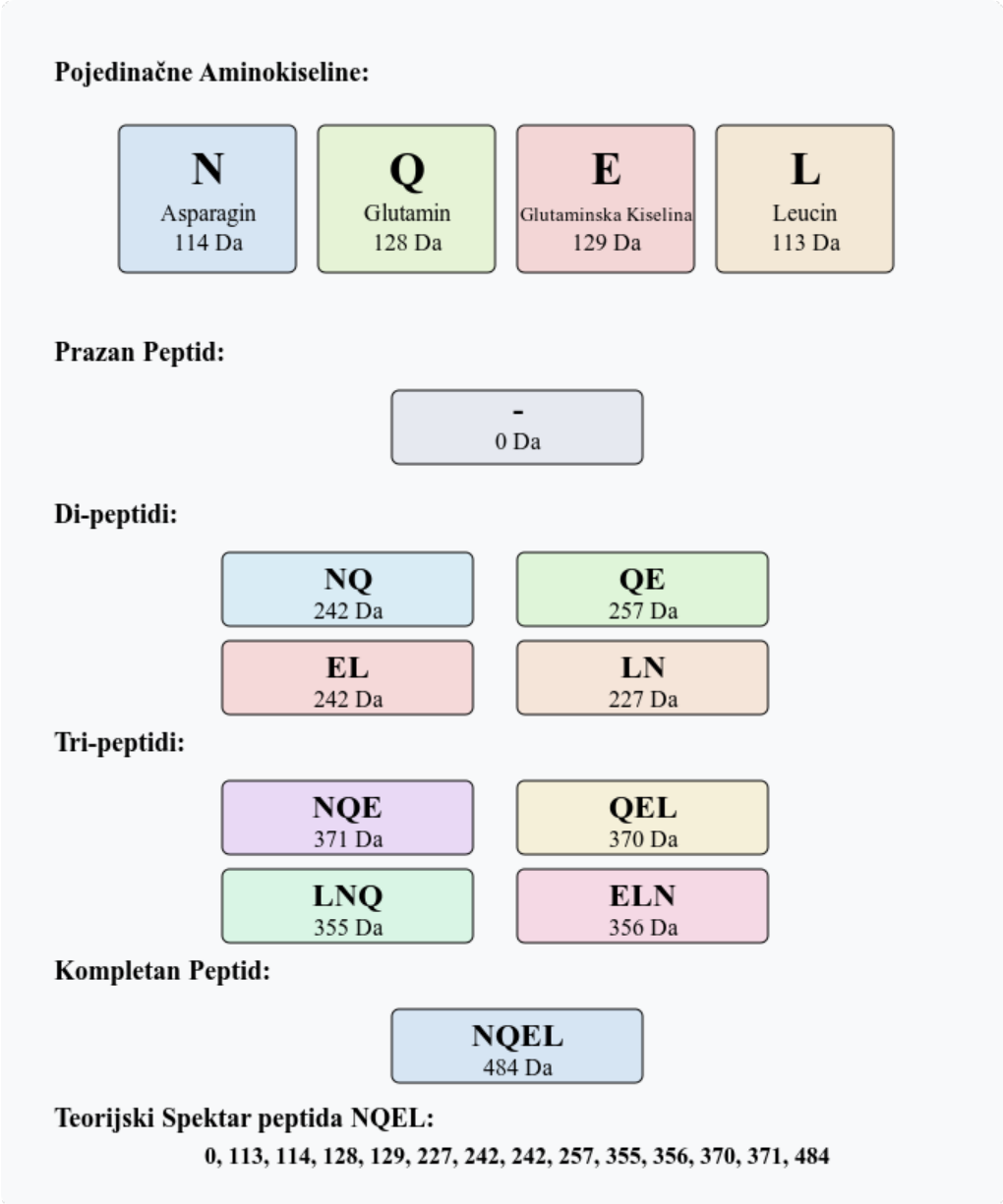
```
 V    K     L     F     P    W     F     N     Q     Y
99 + 128 + 113 + 147 + 97 + 186 + 147 + 114 + 128 + 163 = 1322
```

2.4 Teorijski spektar peptida

Teorijski spektar peptida predstavlja mase svih mogućih podpeptida, uključujući 0 i masu celog peptida. Na osnovu peptida možemo lako da odredimo teorijski spektar ali na osnovu spektra ne možemo lako da odredimo koji je peptid u pitanju.

Problem sekvenciranja ciklopeptida samim tim se svodi na problem kako rekonstruisati ciklični peptid na osnovu njegovog teorijskog spektra. U nastavku će biti prikazani nekoliko različitih algoritama.

Kao ulaz u svaki od ovih algoritama očekuje se eksperimentalni spektar, odnosno spektar koji je dobijen uz pomoć masenog spektrometra za neki peptid. Na Slici 2.4 su prikazane mase svih podpeptida peptida NQEL koje se dobijaju uz pomoć masenog spektrometra, kao i masa praznog peptida i celog peptida, takođe je prikazan i teorijski spektar.



Slika 2.4: Teorijski spektar peptida NQEL koji prikazuje sve moguće podpeptide, njihove mase i njegov teorijski spektar

Glava 3

Algoritmi za sekvenciranje

U ovom odeljku biće prikazani algoritmi za određivanje cikličnog peptida na osnovu poznatog eksperimentalnog spektra. Neki od algoritama koji će biti objašnjeni su:

- **Algoritam grube sile** (*Brute force*): Direktan pristup gde se isprobavaju sve moguće kombinacije da bi se našlo optimalno rešenje.
- ***Branch and Bound***: Optimizovan algoritam koji će odbacivati kandidate čim prestanu da budu potencijalno rešenje.
- ***Leaderboard* algoritam**: Algoritam koji održava listu N najboljih kandidata za rešenje i na osnovu njih smanjuje broj potencijalnih kandidata.
- **Spektralna konvolucija**: Određuje aminokiselinaekoje mogu da učestvuju u peptidu na osnovu eksperimentalnog spektra.
- ***DeepNovo***: Metoda zasnovana na dubokom učenju koja omogućava sekvenciranje peptida bez oslanjanja na baze podataka.

3.1 Gruba sila (Brute Force)

Najosnovniji pristup sekvenciranju koji sistematski ispituje sve moguće kombinacije aminokiselina dok ne pronađe sekvencu koja najbolje odgovara eksperimentalnom spektru. Iako jednostavan za implementaciju, ovaj pristup postaje neizvodiv za duže sekvence zbog eksponencijalnog rasta prostora pretrage.

Na primer, za peptid mase 579 Da, algoritam će generisati sve moguće kombinacije aminokiselina i proveriti da li njihova ukupna masa odgovara zadatoj masi. Kao

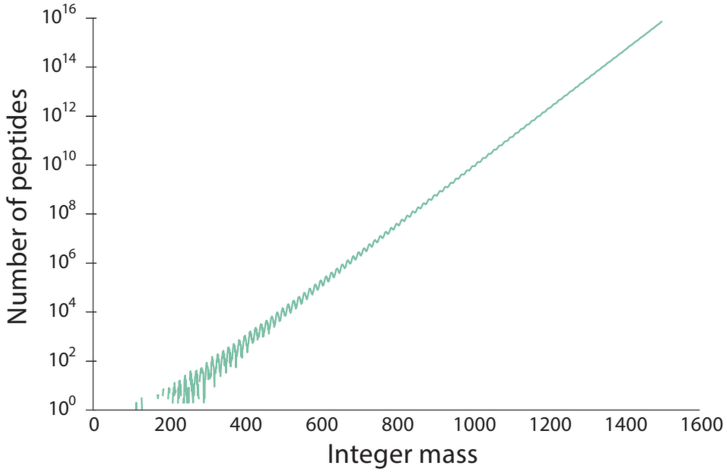
što je prikazano u Tabeli 3.1, mogu postojati različiti peptidi sa istom ukupnom masom (**FPAYT** i **QNWGS**), što dodatno komplikuje problem.

F	147 Da	**Q**	128 Da
P	97 Da	**N**	114 Da
A	71 Da	**W**	186 Da
Y	163 Da	**G**	57 Da
T	101 Da	**S**	94 Da
Ukupna masa: 579 Da		**Ukupna masa:** 579 Da	

Tabela 3.1: Poređenje aminokiselina i njihove mase

Da bi se utvrdilo koji od peptida je tačno rešenje, algoritam mora da generiše teorijski spektar za svaki kandidat peptid i uporedi ga sa eksperimentalnim spektrom. Ovo dodatno povećava računsku složenost algoritma, ali je neophodno za pronalaženje tačnog rešenja.

Na slici 3.1, možemo da vidimo koliko postoji različitih peptida za istu masu, samim tim možemo zaključiti da je izvršavanje algoritma grube sile veoma neefikasno.



Slika 3.1: Broj peptida koji imaju istu masu

Na osnovu prethodnog teksta definiše se pseudokod za algoritam grube sile koji se može videti kao Algoritam 1.

Algoritam 1: Gruba sila

```
Funkcija GrubaSila(eksperimentalniSpektar)
    peptidi ← Lista sa praznim stringom
    rezultati ← Prazna lista
    ciljna_masa ← Poslednji element eksperimentalniSpektar
    while peptidi ≠ Prazno do
        prosireni ← Proširi(peptidi)
        kandidati ← Prazna lista
        foreach peptid ∈ prosireni do
            masa ← IzračunajMasu(peptid)
            if masa = ciljna_masa then
                if CikličniSpektar(peptid) = eksperimentalniSpektar then
                    Dodaj peptid u rezultati
                end
            else
                if masa < ciljna_masa then
                    Dodaj peptid u kandidati
                end
            end
        end
        peptidi ← kandidati
    end
    return rezultati
```

Objašnjenje pomoćnih funkcija:

- **Proširi(peptidi)** – sve preostale peptide proširuje sa svakom mogućom aminokiselinom i vraća proširenu listu.
- **IzračunajMasu(peptid)** – računa ukupnu masu peptida sabiranjem masa svih aminokiselina koje čine taj peptid.
- **CikličniSpektar(peptid)** – za peptid koji je potencijalno rešenje generiše se ciklični spektar koji se poredi sa zadatim eksperimentalnim spektrom.

3.2 Branch and Bound

3.3 Leaderboard algoritam

Algoritam koji održava listu (leaderboard) najboljih kandidata tokom pretrage. U svakoj iteraciji, algoritam proširuje sekvence sa svim mogućim aminokiselinama i zadržava samo najbolje kandidate za dalju obradu.

3.4 Spektralna konvolucija

3.5 DeepNovo

Glava 4

Elektronska platforma

4.1 Arhitektura sistema

Backend (Django)

Frontend (Next.js)

4.2 Funkcionalnosti platforme

4.3 Korišćenje platforme

Glava 5

Zaključak

Razvijena elektronska platforma za sekvenciranje antibiotika predstavlja moćan edukativni alat koji objedinjuje teorijske koncepte sa praktičnom implementacijom. Kroz interaktivne simulacije i vizuelizacije, platforma omogućava dubinsko razumijevanje kompleksnih algoritama sekvenciranja.

Glavni doprinosi ovog rada uključuju:

- Integraciju više algoritama za sekvenciranje u jedinstvenu platformu
- Vizuelizaciju koraka algoritama u realnom vremenu
- Kreiranje interaktivnog okruženja za učenje
- Poboljšanje dostupnosti bioinformatičkih alata studentima

Budući radovi mogu se usredsrediti na proširenje platforme sa dodatnim algoritmima, poboljšanje performansi postojećih implementacija i integraciju sa stvarnim eksperimentalnim podacima iz masene spektrometrije.

Bibliografija

Biografija autora

Miloš Milaković rođen je 6. avgusta 1998. godine u Beogradu. Smer Informatika na Matematičkom fakultetu Univerziteta u Beogradu upisao je 2017. godine, a završio 2021. godine sa prosečnom ocenom 8.54. Nakon toga je upisao master studije na istom smeru. Od septembra 2021. do marta 2024. godine je zaposlen na poziciji *Software* developer u firmi **Endava**. Od marta 2024. godine do sada je zaposlen na poziciji *Software* developer u firmi **LotusFlare**. Projekti na kojima je radio su uglavnom bili zasnovani na veb tehnologijama (*Telco* industrija i *FinTech* industrija), a osnovni programski jezik u kojem su projekti rađeni su *Python* i *Lua*.